Detecting Significant Behaviour in Tweets using Machine Learning

Faisal Shahzad
Faculty of Computing,
The Islamia University of Bahawalpur.
Bahawalpur, Pakistan.
fs.ahmad65@gmail.com

Muhammad Asad Ullah
Department of Information Technology,
Faculty of Computing,
The Islamia University of Bahawalpur.
Bahawalpur, Pakistan
muhammad.asadullah@iub.edu.pk

Muhammad Adnan Khan
Department of Computing,
Skyline University College.
Sharjah,
United Arab Emirates.
muhammad.adnan@skylineuniversity.ac.ae

Nouh Sabri Elmitwally
School of Computing and Digital Technology,
Birmingham City University, Birmingham
B4 7XG, United Kingdom.
nouh.elmitwally@bcu.ac.uk

***Abstract*—Sentiment Analysis is a crucial area of study within the realm of Computer Science. With the rapid advancement of Information Technology and the prevalence of social media, a substantial volume of textual comments has emerged on web platforms and social networks such as Twitter. Consequently, individuals have become increasingly active in disseminating both general and politically-related information, making it imperative to examine public responses. Many researchers have harnessed the unique features and content of social media to assess and forecast public sentiment regarding political events. This study presents an analytical investigation employing data from general discussions on Twitter to decipher public sentiment regarding the crisis in Pakistan. It involves the analysis of tweets authored by various ethnic groups and influential figures using Machine Learning techniques like the Support Vector Classifier (SVC), Decision Tree (DT), Naïve Bayes (NB) and Logistic Regression. Ultimately, a comparative assessment is conducted based on the outcomes obtained from different models in the experiments.**

Keywords—hate speech, sentiment analysis, tweets, political opinion, insert.

I. INTRODUCTION

. The rapid expansion of online social networks (OSNs) has made communication platforms in high demand. This trend has facilitated broader data sharing, exploration, and information sharing, all without being limited by geographic boundaries (Antypas, Preece, and Camacho-Collados 2023). The amount of content generated through social media channels, especially Twitter, is staggering. Twitter serves as an online environment for information and social interaction, where users communicate through short tweets (Lagman et al. 2018). It has become the leading social media platform, with millions of users posting tweets regularly.

The volume of public opinion data has increased exponentially(De Choudhury et al. 2016). The ability to identify these perspectives on political events and issues is critical to shaping international agreements, policies and standards. Officials rely on these sentiments to inform their decisions, making it imperative that they closely monitor these data for future policy decisions(Davidson et al. 2017).. Polls have traditionally served as the primary means of gathering public opinion, but they often present a number of challenges. These studies struggle to provide nuanced analysis or to reveal the underlying motivations, subjectivity, and intentions behind public sentiment. These limitations make opinion polls unreliable, highlighting the need for more sophisticated methods of understanding public opinion. The advent of social media, with its large user base, diverse topics, and large number of user-generated content, has emerged as an important tool for predicting human sentiment (Chung and Mustafaraj 2011). Using advanced techniques to elicit political opinion on these platforms could provide a faster, more accurate, and more cost-effective alternative to traditional polls(Liu 2011).

Although many studies have examined the potential of social media mining to analyze and predict political opinion, most of them have been event-oriented and used specialized techniques. Furthermore, most of these studies rely on sentiment analysis to assess consumer sentiments rather than political stances. Analysis of sentiment in politics often relies on personal consistency, ignoring the differences in consumer opinion(Alhojely 2016). As a result, a sentiment analyst can classify a tweet, phrase, comment, or expression as “positive”, “negative” or “neutral.” Moreover, only a small number of researchers have statistically analyzed the current political climate, rather than making predictions about public sentiment.

This article presents an analytical and comparative study of a Twitter text dataset to assess public attitudes in different countries regarding the Pakistan crisis which involves inflation, terrorism, establishment involvement in politics and political affairs related to Pakistan. The data used in this study date from 2018-2023. The proposed approach uses machine learning models, including Support Vector Classifier (SVC), Decision Tree (DT), Naïve Bayes (NB) and Logistic Regression algorithms, and the results are given a comparative analysis of these examples. Section 2 provides an overview of previous research in this area. Section 3 briefly describes the methodology used in this research. Section 4 describes in detail the use of the research. The conclusions of the analysis and evaluations are presented in section 5, while section 6 discusses the conclusions and future perspectives.

TABLE 1 TWITTER MOST USER COUNTRIES STATISTICS

Countries	Twitter Users 2021 (in millions)
United States	69.3
Japan	50.9
India	17.5
United Kingdom	16.45
Brazil	16.2
Indonesia	14.05
Turkey	13.6
Saudi Arabia	12.45
Mexico	11
France	8
Philippines	7.85
Spain	7.5
Thailand	7.35

II. DATASET

A. Data Collection

Numerous Twitter-based datasets are available online, either for free or for purchase, to understand public sentiment regarding social or political matters. However, there is no dataset available to evaluate public behavior towards Pakistan. Consequently, we collected fresh, relevant Twitter data related to this issue. To achieve this, we combined various existing datasets and filtered out important tweets to create a new dataset that met our requirements. The new dataset was saved in a CSV (comma separated values) file format to enable easy analysis. We used several Twitter trends, such as #pakistan, #jihad #islamicstate, #islam, #terrorist, #terrorism, #NawazSharif, and #ImranKhan, to obtain data. The table given below shows the top most highest hashtags used on twitter with respect to Pakistan.

TABLE 2. TOP HASHTAGS WITH RESPECT TO PAKISTAN

Trending Hashtag / Topic	Tweet Count
#عمران_ہی_تحریک_انصاف_ہے	102209
#قاسم_کے_ابا	81575
#PMShehbazinTurkiye	13636
#غدار_فتنے_نہیں_چھوڑیں_گے	52415
#May9th_FalseFlag	121529
Gundogan	133054
Benzema	113572
pakistan and turkey	below 10K
Ben Stokes	below 10K
De Gea	12305
Drop a Picture With a Mask	below 10K
Imam Khomeini	below 10K
Sancho	59109
Usman Buzdar	below 10K

The data obtained through these trends contained only tweets that exhibited behavior linked to Pakistan. The newly created dataset comprised 7013 tweets, and we applied different classification algorithms to it.

* Corresponding author Muhammad Adnan Khan

TABLE 3 TOP MOST TOPIC FOR DATA COLLECTION

English Topics	Urdu Topics
PTI	جنرل قمر جاوید باجوہ
General Qamar Javed Bajwa	جہاد
Jihad	نواز شریف
Nawaz Sharif	عمران خان
Imran Khan	اسلاموفوبیا
Islamophobia	تحریک لبیک پاکستان
TLP	امپورٹڈ حکومت
Imported Govt	دہشت گرد
Terrorism	کشمیر
Kashmir	پاکستان تحریک انصاف

B. Data Labeling

Data labeling is vital in supervised learning, and its accuracy significantly impacts model performance. While manual and automatic labeling methods are used depending on the data type, labeling criteria, and available resources, semi-supervised approaches can help reduce labeling costs while still achieving accurate classification results(Zhang, Jafari, and Nagarkar 2021). It is essential to ensure labeling consistency and accuracy, provide sufficient training to annotators, and have quality control mechanisms in place.

III. TECHNIQUES

The proposed techniques for sentiment analysis are briefly discussed below. The techniques are:

A. Support Vector Machine (SVM)

Support Vector Classification (SVC) is utilized to address classification tasks. In this supervised algorithm, data is partitioned using the optimal decision boundary known as the hyperplane. This hyperplane is established with the aid of crucial data points, referred to as support vectors. Furthermore, SVC involves both a positive hyperplane, which intersects with one or more of the nearest positive attributes, and a negative hyperplane, which intersects with one or more of the nearest negative points. The ideal hyperplane is the one that maximizes the margin, which represents the space between the positive and negative hyperplanes.

B. Decision Tree

Decision trees, abbreviated as DT, play an important role in solving classification and regression problems. This supervised learning method uses a tree structure to convey predictions generated through a series of element-oriented partitions. Starting from the first node and ending with the definition made in leaf nodes, it includes basic concepts such as root nodes, leaf nodes, decision nodes, and pruning and subtrees. This approach uses a series of conditional tests that work in the same way as a set of if-else statements. It moves to the next node associated with the solution if the condition is confirmed to be true.

C. Näive Bayes

Naïve Bayes is a supervised machine learning algorithm that relies on Bayes' Theorem and is commonly employed for addressing classification tasks. Its primary application lies in the realm of text classification, particularly when dealing with large, high-dimensional training datasets. This

algorithm is known for its ability to construct efficient and precise machine learning models, facilitating rapid predictions.

$$P(Y|X)=\frac{P(X|Y)P(Y)}{\Sigma_{Y^T=1}^{C} P\left(X\middle|Y^{'}\right)P(Y)} \quad (1)$$

- The variables X and Y represent the instance of a sentiment label and the corresponding sentiment class (weak positive, moderate positive, strong positive, weak negative, moderate negative, strong negative, or neutral), respectively.
- P(X|Y) refers to the probability of an instance occurring in a particular class for each value of Y, also known as the class-conditional density.
- P(Y) represents the prior probability of a class.

D. Feature Extraction using TF/IDF

When conducting sentiment analysis on text, various features are typically used. These features can be derived from different levels, such as word-level, character level, or even include POS tags, hashtags, emoticons, user tags, and abbreviations(Bounabi, El Moutaouakil, and Satori 2017). For instance, a word-level feature could be the term "airstrike" or "terrorist," while a word-level n-gram feature could be "doing_good" or "good_job." On the other hand, a character level n-gram feature could be "b," "be," "beh," "av," "ave," "beha," or "behave." Moreover, word clusters such as "maybe," "probably," and "prob" may be collapsed into the same cluster. The feature building phase aims to convert text data into a manageable representation that can be understood by the algorithm(Shang and Ran 2022). To achieve this, a feature vector is constructed, typically using a weighting scheme like Term Frequency-Inverse Document Frequency (TF-IDF)(Wu et al. 2008).

$$TF=\frac{\text{number of times the term t appears in document d}}{\text{total number of terms in document d}} \quad (2)$$

We considered *d* as tweets and *t* as occurrence of word in tweet.

$$IDF=\frac{\text{total tweets in the dataset}}{\text{tweets that contain the term t}} \quad (3)$$

$$TF\text{-}IDF(t,d)=TF(t,d)\times IDF(t,d) \quad (4)$$

E. Bag of Words (BoW)

The Bag of Words (BoW) method is a method for representing textual information in machine learning models. This method converts any text into fixed length vectors by counting the number of each word. This feature, called vectorization, is often used to solve problems such as language modeling and text segmentation by capturing key features of text. BoW offers great flexibility in adapting to specific text datasets.

IV. IMPLEMENTATION

This section outlines the typical procedure for conducting sentiment analysis. Figure 1 illustrates a schematic representation of the suggested approach. A brief explanation of the process follows:

A. Text Preprocessing

Tokenization is a fundamental text preprocessing technique used in machine learning to break down textual data into smaller meaningful units called tokens(Xiaofeng, Wei, and Aiping 2020). In this process, a large text corpus is divided into smaller subunits called tokens, which could be words, phrases, or sentences. These tokens can be further processed and used as input for machine learning models(Grefenstette 1999; Roth et al. 2021). This involves transforming characters into a consistent case, removing punctuation marks, special characters, and other non-alphanumeric characters, as well as converting numbers and dates into a standard format(Chen and Ku 2002). For example, suppose we have a dataset containing text data with different capitalization and punctuation.

In Natural Language Processing (NLP), stop words are commonly used words that are considered to be insignificant in the context of text analysis. Examples of stop words include articles (e.g., "the," "an," "a"), conjunctions (e.g., "and," "but," "or"), and prepositions (e.g., "in," "on," "at"). Stop words removal is the process of eliminating these words from a text corpus to reduce the noise in the data and focus on the more relevant words(Sarica and Luo 2021). In the case of tweets, stop words removal is particularly useful because tweets have a limited length of 280 characters, and the presence of stop words can increase the noise-to-signal ratio. The goal of this process is to improve the quality of the data used for training machine learning models. In the context of tweet data, irregular terms may include hashtags, user mentions, emoticons, slang, abbreviations, and other non-standard or informal language(Kharde and Sonawane 2016). For example, consider the following tweet: "Just had the best pizza #yum #pizzalove @pizzahut 💸❤️". This tweet contains several irregular terms, such as the hashtags #yum and #pizzalove, the user mention @pizzahut, and the emoticons 💸 and ❤️.

S is commonly used in natural language processing (NLP) and machine learning (ML) to improve text analysis and classification by reducing the number of variations of words(Popovič and Willett 1992). For example, the words "running," "runner," and "runners" all share the same stem, which is "run".

- Porter stemming is one of the most commonly used algorithms in NLP and ML. It applies a set of rules to remove common English suffixes and prefixes from words(Singh and Gupta 2016).
- Snowball stemming is an extension of the Porter stemming algorithm and provides additional rules for stemming words in different languages, such as French, German, and Spanish(Moral et al. 2014).

Lemmatization is a technique for normalizing text that involves identifying the basic form or lemma of each inflected word in a document or request(Balakrishnan and Lloyd-Yemoh 2014). The benefits of lemmatization are similar to those of stemming, as it can help to reduce ambiguity and precision problems caused by inflectional word forms(Korenius et al. 2004). This can result in issues with accuracy and completeness(Alkula 2001).

When conducting sentiment analysis on text, various features are typically used. These features can be derived from different levels, such as word-level, character level, or even include POS tags, hashtags, emoticons, user tags, and abbreviations(Bounabi et al. 2017). For instance, a word-level feature could be the term "airstrike" or "terrorist," while a word-level n-gram feature could be "doing_good" or "good_job." On the other hand, a character level n-gram feature could be "b," "be," "beh," "av," "ave," "beha," or "behave." Moreover, word clusters such as "maybe," "probably," and "prob" may be collapsed into the same cluster. The feature building phase aims to convert text data into a manageable representation that can be understood by the algorithm(Shang and Ran 2022). To achieve this, a feature vector is constructed, typically using a weighting scheme like Term Frequency-Inverse Document Frequency (TF-IDF)(Wu et al. 2008).

B. Lexical sentiment analysis

This procedure entails evaluating the emotional tone of a text through an examination of the meaning and connections of words and phrases. The assessment was carried out utilizing the Python TextBlob library, which offers two key measurements: polarity and subjectivity. Polarity scores span from -1 to 1, where -1 denotes a pessimistic tone, 0 signifies a neutral tone, and +1 indicates an optimistic tone. Meanwhile, subjectivity scores range from 0 to 1 and aid in ascertaining whether a text presents factual data or expresses personal viewpoints.

C. Evaluation

a) Embedding Words

Embedding is a technique used to learn the semantic essence of a language. This process involves converting individual words into vector representations using the CountVectorizer tool from the scikit-learn library.

b) Splitting Dataset

In this step the dataset is splitted into two sets training and testing. The training and testing portion is 80% and 20% respectively.

c) Training the Model

This step involve the implementation of the different machine learning classifiers like support vector classifier, decision tree, naïve Bayes.

V. Experimental Results

TABLE 4 PERFORMANCE TABLE OF ML CLASSIFIRES

Types of polarity	Precision	Recall	F1-score	Accuracy
SVC	0.76	0.76	0.75	76.67%
DT	0.77	0.75	0.74	81%
NB	0.79	0.74	0.75	84.15 %
Logistic Regression	0.72	0.71	0.70	71%

References

Alhojely, Suad. 2016. *Sentiment Analysis and Opinion Mining: A Survey*. Vol. 150.

Alkula, Riitta. 2001. “From Plain Character Strings to Meaningful Words: Producing Better Full Text Databases for Inflectional and Compounding Languages with Morphological Analysis Software.” *Information Retrieval* 4(3–4):195.

Antypas, Dimosthenis, Alun Preece, and Jose Camacho-Collados. 2023. “Negativity Spreads Faster: A Large-Scale Multilingual Twitter Analysis on the Role of Sentiment in Political Communication.” *Online Social Networks and Media* 33. doi: 10.1016/j.osnem.2023.100242.

Balakrishnan, Vimala, and Ethel Lloyd-Yemoh. 2014. “Stemming and Lemmatization: A Comparison of Retrieval Performances.”

Bounabi, M., K. El Moutaouakil, and Kh. Satori. 2017. “A Comparison of Text Classification Methods Method of Weighted Terms Selected by Different Stemming Techniques.” in *Proceedings of the 2nd International Conference on Big Data, Cloud and Applications*, *BDCA’17*. New York, NY, USA: Association for Computing Machinery.

Chen, Hsin-Hsi, and Lun-Wei Ku. 2002. “An NLP & IR Approach to Topic Detection.” *Topic Detection and Tracking: Event-Based Information Organization* 243–64.

De Choudhury, Munmun, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. 2016. *Social Media Participation in an Activist Movement for Racial Equality*.

Chung, Jessica, and Eni Mustafaraj. 2011. “Can Collective Sentiment Expressed on Twitter Predict Political Elections?” *Proceedings of the National Conference on Artificial Intelligence* 2:1770–71. doi: 10.1609/AAAI.V25I1.8065.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. *Automated Hate Speech Detection and the Problem of Offensive Language*.

Grefenstette, Gregory. 1999. “Tokenization.” *Syntactic Wordclass Tagging* 117–33.

Kharde, Vishal A., and S. S. Sonawane. 2016. *Sentiment Analysis of Twitter Data: A Survey of Techniques*. Vol. 139.

Korenius, Tuomo, Jorma Laurikkala, Kalervo Järvelin, and Martti Juhola. 2004. “Stemming and Lemmatization in the Clustering of Finnish Text Documents.” Pp. 625–633 in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, *CIKM ’04*. New York, NY, USA: Association for Computing Machinery.

Lagman, Ace C., Melvin A. Ballera, Jennifer O. Contreras, and Jennalyn G. Raviz. 2018. “Development of Converted Deterministic Finite Automaton of Decision Tree Rules of Student Graduation and Adaptive Learning Environment.” Pp. 267–71 in *ACM International Conference Proceeding Series*. Association for Computing Machinery.

Liu, Bing. 2011. “Opinion Mining and Sentiment Analysis.” Pp. 459–526 in *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, edited by B. Liu. Berlin, Heidelberg: Springer Berlin Heidelberg.

Moral, Cristian, Angélica de Antonio, Ricardo Imbert, and Jaime Ramírez. 2014. “A Survey of Stemming Algorithms in Information Retrieval.” *Information Research: An International Electronic Journal* 19(1):n1.

Popovič, Mirko, and Peter Willett. 1992. “The Effectiveness of Stemming for Natural-language Access to Slovene Textual Data.” *Journal of the American Society for Information Science* 43(5):384–90.

Roth, Tom, Yansong Gao, Alsharif Abuadbba, Surya Nepal, and Wei Liu. 2021. “Token-Modification Adversarial Attacks for Natural Language Processing: A Survey.” *ArXiv*

Preprint ArXiv:2103.00676.

Sarica, Serhad, and Jianxi Luo. 2021. “Stopwords in Technical Language Processing.” *Plos One* 16(8):e0254937.

Shang, Fengjun, and Chunfu Ran. 2022. “An Entity Recognition Model Based on Deep Learning Fusion of Text Feature.” *Information Processing and Management* 59(2). doi: 10.1016/j.ipm.2021.102841.

Singh, Jasmeet, and Vishal Gupta. 2016. “Text Stemming: Approaches, Applications, and Challenges.” *ACM Computing Surveys (CSUR)* 49(3):1–46.

Wu, Ho Chung, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. “Interpreting TF-IDF Term Weights as Making Relevance Decisions.” *ACM Transactions on Information Systems* 26(3):1–37. doi: 10.1145/1361684.1361686.

Xiaofeng, Mu, Wang Wei, and Xu Aiping. 2020. “Incorporating Token-Level Dictionary Feature into Neural Model for Named Entity Recognition.” *Neurocomputing* 375:43–50. doi: 10.1016/j.neucom.2019.09.005.

Zhang, Shikun, Omid Jafari, and Parth Nagarkar. 2021. “A Survey on Machine Learning Techniques for Auto Labeling of Video, Audio, and Text Data.”